Attention enabled multi-agent DRL for decentralized volt-VAR control of active distribution system using PV inverters and SVCs

Cao, Di; Zhao, Junbo; Hu, Weihao; Ding, Fei; Huang, Qi; Chen, Zhe

Published in:
IEEE Transactions on Sustainable Energy

DOI (link to publication from Publisher):
10.1109/TSTE.2021.3057090

Publication date:
2021

Document Version
Accepted author manuscript, peer reviewed version

Attention Enabled Multi-agent DRL for Decentralized Volt-VAR Control of Active Distribution System Using PV Inverters and SVCs

Di Cao, *Student Member, IEEE,* Junbo Zhao, *Senior Member, IEEE,* Weihao Hu, *Senior Member, IEEE,* Fei Ding, *Senior Member, IEEE,* Qi Huang, *Senior Member, IEEE,* Zhe Chen, *Fellow, IEEE*

***Abstract*—This paper proposes attention enabled multi-agent deep reinforcement learning (MADRL) framework for active distribution network decentralized Volt-VAR control. Using the unsupervised clustering, the whole distribution system can be decomposed into several sub-networks according to the voltage and reactive power sensitivity relationships. Then, the distributed control problem of each sub-network is modeled as Markov games and solved by the improved MADRL algorithm, where each sub-network is modeled as an adaptive agent. An attention mechanism is developed to help each agent focus on specific information that is mostly related to the reward. All agents are centrally trained offline to learn the optimal coordinated Volt-VAR control strategy and executed in a decentralized manner to make online decisions with only local information. Compared with other distributed control approaches, the proposed method can effectively deal with uncertainties, achieve fast decision makings, and significantly reduce the communication requirements. Comparison results with model-based and other data-driven methods on IEEE 33-bus and 123-bus systems demonstrate the benefits of the proposed approach.**

***Index Terms*—Voltage regulation, network partition, multi-agent deep reinforcement learning, distribution network, PV inverters, distribution system optimization.**

I. Introduction

The utilization of renewable energy is of great significance to alleviate the current energy and environmental concerns [1]-[2]. However, due to the uncertainties and volatility of renewable energy, its higher-level integration brings numerous technical challenges to the active distribution network (ADN) operations. Among them, the overvoltage issues due to the reverse power flow caused by the increased penetration of renewable energy are of particular attention.

Various approaches have been proposed for ADN voltage regulation. From the perspective of the control framework, they can be divided into three main categories: centralized, distributed autonomous, and distributed coordination control. The centralized control applies optimization algorithms to solve a centralized voltage regulation problem based on the operation information of the whole system [13]. This is challenging since the optimal power flow (OPF) is a non-convex optimization problem. To this end, heuristic methods [3], approximated approaches [4], non-linear optimization [5], and convexification methods [6] have been developed. To further address the uncertainties of renewable energy generations, stochastic programming (SP)-based approaches are developed [7]. However, the SP-based ones have a heavy computational burden as many scenarios need to be considered. Also, SP requires the accurate distribution of random variables, which may not be possible in practice. Different from SP, robust optimization (RO) methods deal with the uncertainty by constructing an uncertainty set and obtain the solutions under the worst scenarios [8-10]. As a result, the outcomes are typically conservative. Model predictive control [11-12] is another way of addressing voltage regulation but has difficulties for large-scale systems. Centralized control needs a central controller based on global information that is challenging given today's limited communication capability in distribution systems. The communication delay can degrade the control performance. It also has disadvantages such as privacy concerns and vulnerability to cyber and physical attacks.

The distributed autonomous control strategies make decisions for voltage regulation based on local observations [14-15]. They are easy to implement but has the problem of finding the global optimal solution due to the lack of cooperation between various control subjects. By contrast, the distributed cooperative control can achieve the coordination between different units with limited communication links [16]. Among them, the partition-based distributed coordination control is attractive in recent years [10], [17-19]. The main idea is to apply a clustering algorithm to partition the whole network into several sub-networks according to the predefined electrical distance. Then the optimization methods are applied to achieve distributed voltage regulation of each cluster. These methods [10], [17-19] need to pre-determine an optimal solution to deal

This work was supported by the National Key Research and Development Program of China (2018YFE0127600). Corresponding author: Weihao Hu.

Di Cao, Weihao Hu are with the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China (e-mail: caodi@std.uestc.edu.cn; whu@uestc.edu.cn)

J. Zhao is with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS 39762 (e-mail: junbo@ece.msstate.edu).

F. Ding is with the Power systems Engineering Center, National Renewable Energy Laboratory, Golden, United States (e-mail: fei.ding@nrel.gov)

Qi Huang is with the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China. He is also with the College of Nuclear Technology and Automation Engineering, Chengdu University of Technology, Chengdu, China. (e-mail: hwong@uestc.edu.cn)

Zhe Chen is with the Department of Energy Technology, Aalborg University, Aalborg, Denmark (e-mail: zch@et.aau.dk).

with the uncertainties of photovoltaic (PV) outputs and load demand. However, the PV outputs can significantly fluctuate when radiation changes fast. According to [20], the PV output may vary by 15% of its rated power in less than one minute. Under this condition, more frequent operations of multiple devices are needed to provide flexible responses to system uncertainties. However, they need to resolve the optimization problem when a new situation is encountered, leading to high computational burdens. Additionally, these methods are model-based and require accurate parameters of the ADN, which is difficult to get [21-22].

To mitigate the model quality issue, machine learning (ML)-based voltage regulation methods for ADN are developed. ML methods can extract knowledge from historical data to deal with system uncertainties. The extracted knowledge has generalizability to new situations without resolving the problem [23-24]. Among them, deep reinforcement learning (DRL) can learn optimal control strategies from data and is suitable for the control and optimization problems [25], [26]. The relationships between the states and actions are learned by continuous interactions with the environment, thus reducing the dependence on the knowledge of system parameters. These DRL algorithms are still based on the centralized control framework and thus subject to the aforementioned issues. To this end, various ML-based decentralized control approaches have been proposed [27-30]. A Q-learning based multi-agent control method is proposed for reactive power dispatch in [27]. [28] develops a multi-agent Q-learning based decentralized control approach for the coordinated regulation of multiple energy storage systems. While in [29], a deep Q-network based multi-agent control framework for the optimization of AND is proposed. Nevertheless, Q-learning algorithms require the discretization of state and action variables. However, most practical power system applications have continuous multi-dimensional state and action space. The naïve discretization of multiple continuous variables may lead to the curse of dimensionality issues. To this end, a multi-agent autonomous voltage control framework based on a multi-agent deep deterministic policy gradient (MADDPG) algorithm is developed for the voltage regulation of transmission systems [30]. It is worth noting that the DDPG algorithm is difficult to stabilize in complicated applications and sensitive to the settings of hyper-parameters. As the complexity of the environment grows exponentially with the number of agents, it is challenging for the MADDPG algorithm to learn a good policy. To address that, this paper develops an attention-enabled MADRL algorithm that adopts centralized training and decentralized execution framework for Volt-VAR control considering the reactive power capability of static var compensation (SVC) and PV inverters. The contributions are:

• The proposed approach can achieve cooperative control of multiple control devices using only local measurement information without a central controller. This is developed based on a novel framework, namely the centralized training, and decentralized execution, where all agents are trained in a centralized manner to learn the coordination control strategy and are executed in a distributed manner to provide near-optimal decisions based on the latest local information. This significantly reduces the communication requirements and avoids the negative impacts on control performance caused by the time delay. Note that most existing distributed control methods still need some communications, which is not the case for our proposed method. This is one of the most important contributions and it distinguishes with existing methods.

• The attention model is integrated with the MADRL method to help each agent attend to the specific information that is mostly related to its reward. This allows maintaining the control performance when the number of agents changes. This distinguishes it from the MADRL algorithm in [30] as it suffers from performance degradation with many agents.

• Compared with other optimization methods, the proposed method can also achieve fast decision makings and effectively deal with violent voltage fluctuations caused by the rapid PV generation. This is because the strategy learned by the DRL method during training can be generalized to new situations without resolving the optimization problem. Furthermore, since only local information is needed, the decision making is rather fast. By contrast, to deal with uncertainties, robust optimization or stochastic optimization methods need to pre-determine a solution, which cannot effectively cope with the violent voltage fluctuations caused by the rapid PV generation changes.

The rest of the paper is organized as follows. In section II, the problem formulation is presented. Section III describes the proposed method. In section IV, the simulation results are illustrated in detail. Finally, Section V concludes this paper.

II. Problem Formulation

Consider an ADN with *N+1* buses, whose lines and buses are denoted as *L* and *N*, respectively. For $i \in \mathrm{N}$, define v_i as the voltage magnitude, and $p_i + jq_i$ as the injected complex power. The injected active power can be divided into the PV generation p_i^g and load demand p_i^l, i.e., $p_i := p_i^g - p_i^l$. The injected reactive power is $q_i := q_i^g + q_i^s - q_i^l$, where q_i^g, q_i^s, and q_i^l represent the reactive power of PV inverter, SVC, and load demand, respectively. Stacking the load demand and generations into vectors, we get $\mathbf{p}^l, \mathbf{q}^l, \mathbf{p}^g, \mathbf{q}^g$ an $\mathbf{q}^s$. Note that OLTCs and capacitor banks may also exist in distribution feeders, but they typically react in a slow timescale. Furthermore, they are usually scheduled offline due to their limited allowable number of daily switches. This makes it difficult for them to address violent voltage fluctuations caused by the rapid changes in PV generations. The SVCs and PV inverters can provide high-speed reactive power support, especially in distribution networks with a high level of PV penetrations. This paper aims to achieve fast voltage control by using SVCs and PV inverters in the scenario of high penetration of DERs. Let π_i denote the parent bus of the non-root bus *i*. $(\pi_i, i) \in \mathrm{L}$ represents the line between the two nodes, which has impedance $r_i + jx_i$. Let $P_i(t) + jQ_i(t)$ the complex power that flows from node π_i to node *i*. Then, the DistFlow equations to model the power flows of the ADN for all buses $i \in \mathrm{N}$ at *t* are given as follows [31]:

$$p_i(t) = \sum_{j \in \mathrm{X}_i} P_j(t) - (P_i(t) - r_i \frac{P_i^2(t) + Q_i^2(t)}{v_{\pi_i}(t)}) \tag{1a}$$

$$q_i(t) = \sum_{j \in \mathrm{X}_i} Q_j(t) - (Q_i(t) - x_i \frac{P_i^2(t) + Q_i^2(t)}{v_{\pi_i}(t)}) \quad (1b)$$

$$v_i(t) = v_{\pi_i}(t) - 2(r_i P_i(t) + x_i Q_i(t)) + (r_i^2 + x_i^2) \frac{P_i^2(t) + Q_i^2(t)}{v_{\pi_i}(t)} \quad (1c)$$

where X_i represents the set of the children nodes. The task aims to minimize the voltage deviation utilizing the reactive power of SVCs and PV inverters. The voltage regulation problem can be described as follows:

$$\underset{\mathbf{q}^s, \mathbf{q}^g}{\text{minimize}} \; \| \mathbf{v}(t) - v_0 \mathbf{1} \|_1 \quad (2)$$

subject to (1a)-(1c)

$$\underline{\mathbf{v}} \le \mathbf{v}(t) \le \overline{\mathbf{v}} \quad (3)$$

$$\underline{q}_i^s \le q_i^s(t) \le \overline{q}_i^s, \quad \forall i \in \mathrm{N}_s \quad (4)$$

$$| q_i^g(t) | \le \sqrt{(s_i^2) - (p_i^g(t))^2}, \quad \forall i \in \mathrm{N}_g \quad (5)$$

where (2) is the objective function and it aims to minimize the sum of the voltage deviation of each node; $\mathbf{q}^s$ and $\mathbf{q}^g$ are control variables; (3) is the voltage constraint for each node, where $\underline{\mathbf{v}}$ and $\overline{\mathbf{v}}$ are the lower and upper limits; (4) describes the reactive power range of SVC, where $\underline{q}_i^s$ and $\overline{q}_i^s$ are its lower and upper limits; N_s denotes the set of nodes connected with SVCs; From (5), the reactive power of the PV inverter at node *i* depends on the active power of PV at time *t*, where s_i denotes the rated apparent power of PV connected to node *i*; N_g denotes the set of nodes connected with PVs.

Existing centralized volt-var control methods suffer from heavy computational burden [32] and are susceptible to the single point of failure [33]. They also lead to a communication bottleneck since the centralized controller must collect information from the whole system [19]. To this end, this paper proposes a distributed attention based multi-agent twin delayed deep deterministic policy gradient (MATD3) algorithm to deal with them.

III. Proposed Distributed MATD3 Volt-VAR Control

The proposed distributed control method contains three main components, namely i) network partition via the clustering algorithm; ii) formulation of the decomposed sub-networks as Markov games and iii) voltage control optimization via the proposed attention-based MATD3 algorithm.

A. Network Partition

The objective of clustering in this paper is to partition the ADN into several sub-networks to identify the voltage control areas. Via the network partition, the centralized optimization problem is divided into several small problems, such that the communication requirements and negative impacts on control performance caused by the communication delay are reduced and the robustness to cyber and physical attacks is improved.

Spectral clustering, an unsupervised learning method derived from spectral graph theory, is used to search for the optimal partition results of ADN in this paper. Since this paper aims to reduce voltage deviation utilizing reactive power of multi resources, the voltage-reactive power sensitivity matrix is used to represent the electrical distance. The affinity matrix is derived based on the sensitivity matrix via

$$w_{i,j} = w_{j,i} = \sum_{i=1,j=1}^{N} \exp(\frac{-\| x_i - x_j \|^2}{2\sigma^2}) \quad (6)$$

where $w_{i,j}$ is the component of the similarity matrix $\boldsymbol{W}$; x_i represents the *i*th row of the sensitivity matrix; σ is the coefficient that controls the width of the neighborhood. The diagonal degree matrix $\mathbf{D}$ can be obtained by $d_i = \sum_{j=1}^{n} w_{ij}$, where d_i denotes the *i*th diagonal element of *D*. After that, the Laplacian matrix *L* is calculated via $\mathbf{L} = \mathbf{D} - \mathbf{W}$. The clustering problem can be transferred to a graph partition problem and the objective function is as follows [34]:

$$\mathbf{F}_{Ncut}(\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_k) = \frac{1}{2} \sum_{i=1}^{k} \sum_{m \in A_i, n \in \bar{A}_i} \frac{w_{m,n}}{vol(\mathbf{A}_i)} \quad (7)$$

where *k* is the total number of clusters; $\mathbf{A}_i$ denotes the *i*th group of the clustering results; $\bar{\mathbf{A}}_i$ is the complement set of $\mathbf{A}_i$; $vol(\mathbf{A}_i)$ represents the weighted sum of all edges in A_i. The objective of (7) is to maximize the internal similarity of subgraphs. According to [34], the optimization of (7) can be reformulated as:

$$\begin{aligned} &\arg\min tr(\mathbf{F}^T \mathbf{D}^{-1/2} \mathbf{L} \mathbf{D}^{-1/2} \mathbf{F}) \\ &s.t. \;\; \mathbf{F}^T \mathbf{F} = \mathbf{I} \end{aligned} \quad (8)$$

The optimal partition results can be obtained by constructing a space using the eigenvectors corresponding to the first k_1 eigenvalues of the matrix $\mathbf{D}^{-1/2} \mathbf{L} \mathbf{D}^{-1/2}$, and clustering the eigenvectors in the space by the K-means algorithm.

Remark: The key factor that affects the partition result of the distribution system is the network reconfiguration. The topology of the network changes after reconfiguration, leading to the change of the voltage-reactive power sensitivity matrix as well as the partition result. However, the topology changes typically yield local impacts on voltage-reactive power sensitivity. Thus, only the local sensitivity indices corresponding to the topology changes need to be updated, which can be done quickly.

B. Formulation of Markov Games

After network partition, the whole network is divided into serval sub-regions. Then, the decentralized voltage regulation of multi sub-networks is formulated as Markov Games (MGs), a multi-agent extension of the Markov decision process. In the MGs, each sub-network is modeled as an adaptive agent. At each time-step, each agent observes the regional system state $\mathbf{s}_j$, including the active and reactive power of loads, and the active power generation of PVs. Based on the observed information, each agent makes decisions $\mathbf{a}_j$ to schedule the reactive power of SVCs and PV inverters. After all agents' actions are executed, each agent obtains a reward, which represents the total voltage deviation of the whole system. Then, the system transfers to the next state. MGs provide a mathematical representation for modeling the distributed

decision-making process. The key components for an MG include state set S , action set A , and reward function R . They are described as follows:

• S : $\mathbf{s} \in \mathrm{S}$ contains the states for all agents. For agent *j*, the state $\mathbf{s}_j \in \mathbf{s}$ is the local observation of sub-network *j*, including (p_i^l, q_i^l, p_i^g), where *i* is the index of the node in sub-network *j*.

• A : $\mathbf{a} \in \mathrm{A}$ contains the actions for all agents. For agent *j*, the action $\mathbf{a}_j \in \mathbf{a}$ includes (α_i^g, α_i^s). The control variables in (2) can be derived as:

$$q_i^g = \alpha_i^g \sqrt{(s_i)^2 - (p_i^g)^2}, \quad -1 < \alpha_i^g < 1 \tag{9}$$

$$q_i^s = \alpha_i^s \overline{q}_i^s, \quad -1 < \alpha_i^s < 1 \tag{10}$$

• R : $r \in \mathrm{R}$ is the immediate reward the agents obtain after the action $\mathbf{a}$ is executed. In this context, all agents share the same reward: $r = -\| \mathbf{v}(t) - v_0 \mathbf{1} \|_1 + \beta$, where $\| \mathbf{v}(t) - v_0 \mathbf{1} \|$ represents the total voltage deviation of all nodes in the DN; β is the penalty term when voltage constraint (3) is not satisfied.

At each time step, agent *j* makes a decision $\mathbf{a}_j$ based on the local observation $\mathbf{s}_j$ in the sub-network *j*. When all agents complete actions, they obtain a shared reward r, and then the system transfers to the next state. This is an MG and the aim of each agent is to learn a policy, which maps its local observation $\mathbf{s}_j$ to action $\mathbf{a}_j$ so as to maximize the discounted cumulative reward from the current time-step onward, $\sum_{k=t}^{T} \gamma^{k-t} r(k)$, where $\gamma \in [0,1]$ is the discount factor that balances the importance between the future and immediate reward.

Remark: some distribution systems may not have SVCs, DSTATCOM, or fixed/switched capacitor. For a high penetration of PV integration system, SVC may not be necessary if the PV reactive power capability is properly utilized. The proposed method is general enough to address that.

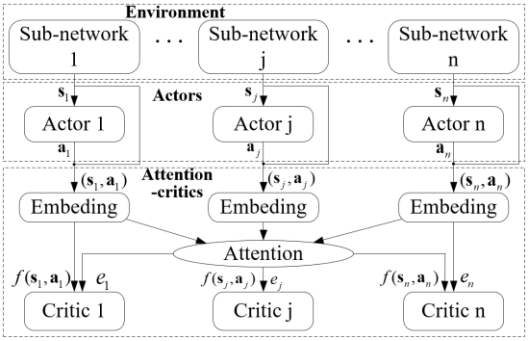


Fig. 1. The architecture of the proposed attention-based MADRL method.

C. Attention Based MATD3 Algorithm for Voltage Control

The proposed attention-based MATD3 algorithm is developed to solve the formulated MGs. MATD3 is one of the MADRL algorithms. To enhance its scalability when dealing with more sub-regions, we improve the original MATD3 algorithm using the attention model. Each sub-network is modeled as a twin delayed deep deterministic policy gradient (TD3) agent, which is composed of the actor and critic networks. The actor maps the local observation $\mathbf{s}_j$ to action $\mathbf{a}_j$ and is the policy function. The critic maps the global information $(\mathbf{s}, \mathbf{a})$ to a scalar, which is a judgment of action $\mathbf{a}_j$ considering the impact on other agents. The coordinated control strategy is achieved by adopting a centralized training framework, among which the actor and critic functions of each agent are trained against each other iteratively until the critic provides better judgment and the actor can make decisions with reduced voltage deviation. To further enhance the capability of MATD3 in dealing with many agents, the attention mechanism is developed. It allows each agent to focus on the specific information that is mostly related to the reward. The architecture of the proposed method is shown in Fig. 1 and the details are elaborated below.

1) Attention-Critic Functions

For agent *j*, the critic function $Q_j(\cdot)$ takes the global state $\mathbf{s}$ and actions of all agents $\mathbf{a}$ as inputs and outputs the action value of the agent *j*. To address the uncertainties of ADN, a deep neural network (DNN) is advocated to approximate the critic functions as follows:

$$Q_j(\mathbf{s}, \mathbf{a}) = g_j(\mathbf{s}, \mathbf{a}) = g_{j,l}[...g_{j,1}(\mathbf{s}, \mathbf{a})] \tag{11}$$

$$g_{j,i} = \sigma(\mathbf{W}_i^c * \mathbf{o}_{i-1} + \mathbf{b}_i^c), \quad i = 2,3,...,l \tag{12}$$

where $g_j(\cdot)$ represents the parameterized critic function of agent *j* approximated by DNN; $g_{j,i}$ represents the function map of the *i*th layer NN; $\mathbf{W}_i^c$ and $\mathbf{b}_i^c$ represent the weight matrix and bias vector of the *i*th layer, respectively; σ represents the activation function; $\mathbf{o}_{i-1}$ is the output of the *(i-1)th* layer. The critic function of agent *j* is parameterized by the parameters of DNN $\mathrm{Q}^{c_j} = \{\mathbf{W}_1^c, \mathbf{b}_1^c, ..., \mathbf{W}_l^c, \mathbf{b}_l^c\}$.

For MADRL, the complexity of the problem increases exponentially with the number of agents. Therefore, it is challenging for the agents to learn good policies when the population is large. To this end, the attention mechanism based critic is developed in this paper. The input of critic function $(\mathbf{s}, \mathbf{a})$ is replaced with $f_j(\mathbf{s}_j, \mathbf{a}_j)$ and e_j , where $f_j(\cdot)$ represents the embedding function of agent *j*, and e_j is the output processed by the attention, representing the weighted sum of other agents' value [35]:

$$e_j = \sum_{i \neq j} \alpha_i \cdot u_i = \sum_{i \neq j} \alpha_i \cdot \mathrm{ReLU}(\mathbf{T} \cdot f_i(\mathbf{s}_i, \mathbf{a}_i)) \tag{13}$$

where ReLU represents the activation function; $\mathbf{T}$ is the linear transformation matrix; α_i represents the attention weight obtained by comparing the similarity between embedding of agent *i*, $f_i(\mathbf{s}_i, \mathbf{a}_i)$ and that of agent *j*, $f_j(\mathbf{s}_j, \mathbf{a}_j)$ using the query-key system [35]:

$$\alpha_i \propto \exp((f_i(\mathbf{s}_i, \mathbf{a}_i))^T \mathbf{W}_k^T \mathbf{W}_q f_j(\mathbf{s}_j, \mathbf{a}_j)) \tag{14}$$

where $\mathbf{W}_k$ and $\mathbf{W}_q$ are the transformation matrices. The calculated similarity value between two embeddings is then passed to a softmax to obtain the attention weight α_t^i. The parameters of the attention model $\mathrm{Q}^a = \{\mathbf{W}_k, \mathbf{W}_q, \mathbf{T}\}$ give rise to

a weighted sum of contributions from all other agents for agent *j*. The parameters of attention-critic Q^{Q_j} include the parameters of critic function Q^{c_j} and those of attention model Q^{a}, which are optimized by minimizing the following loss [36]:

$$L(\mathrm{Q}^{Q_j}) = (Q_j(f_j(\mathbf{s}_j, \mathbf{a}_j), e_j) - y)^2 \quad (15)$$

$$y = r + \gamma Q_j(f_j(\mathbf{s}'_j, \mathbf{a}'_j), e_j)|_{\mathbf{a}'_j = p_j(\mathbf{s}'_j)} \quad (16)$$

where *y* is the target. The critic function is optimized by minimizing the "distance" between $Q_j(\cdot)$ and the target. However, the training process may be unstable since the critic being updated is also used for calculating the target *y*. To solve this problem, the target functions $Q'_j(\cdot)$ and p'_j are introduced. A pair of critics $(Q_{j,1}, Q_{j,2})$ is used for the calculation of target *y* to address the overestimation problem caused by the function approximation error in actor-critic based methods. Then, (16) is rewritten as [37]:

$$y = r_t^j + \gamma \min_{n=1,2} Q'_{j,n}(f_j(\mathbf{s}'_j, \mathbf{a}'_j), e_j)|_{\mathbf{a}'_j = p_j(\mathbf{s}'_j)} \quad (17)$$

where $Q'_{j,n}(\cdot)$ represents the *n*th target critic of agent *j*.

2) Actor Functions

The actor is the policy function that aims at maximizing the output of the critic by making the decision $\mathbf{a}_j$ under the state $\mathbf{s}_j$. DNN is advocated to deal with the nonlinearity when solving the OPF problem, yielding:

$$\mathbf{a}_j = p_j(\mathbf{s}_j) = p_{j,l}[...p_{j,1}(\mathbf{s}_j)] \quad (18)$$

$$p_{j,i} = \sigma(\mathbf{W}_i^a * \mathbf{o}_{i-1} + \mathbf{b}_i^a), \quad i = 2,3,...,l \quad (19)$$

where $p_j(\cdot)$ represents the parameterized policy function of agent *j* approximated by DNN; $p_{j,i}$ represents the function map of the *i*th layer NN; $\mathbf{W}_i^a$ and $\mathbf{b}_i^a$ represent the weight matrix and bias vector of the *i*th layer NN, respectively; σ is the activation function; $\mathbf{o}_{i-1}$ is the output of the *(i-1)*th layer NN. Then, the policy function of agent *j* is parameterized by $\mathrm{Q}^{\mu_j} = \{\mathbf{W}_1^a, \mathbf{b}_1^a, ..., \mathbf{W}_l^a, \mathbf{b}_l^a\}$. They are optimized via the policy gradient [35]:

$$\nabla_{\mathrm{Q}^{\mu_j}} J(\mathrm{Q}^{\mu_j}) = E_{S_t, A_t \sim D}[\nabla_{\mathrm{Q}^{\mu_j}} p_j(\mathbf{a}_j | \mathbf{s}_j) \nabla_{\mathbf{a}_j} Q_j(f_j(\mathbf{s}_j, \mathbf{a}_j), e_j)|_{\mathbf{a}_j = p_j(\mathbf{s}_j)}] \quad (20)$$

3) Replay Buffer Mechanism

Since DNN is utilized to fit the actor and critic functions, the input data for training should be independent and identically distributed. However, the data are highly correlated for the DRL algorithm. To this end, the replay buffer is utilized, where each agent employs memory to store the transitions $(\mathbf{s}_j, \mathbf{a}_j, r, \mathbf{s}'_j)$. The mini-batch experiences are sampled at each time step to calculate the gradient and optimize the parameters of networks. This mechanism helps break the correlation between data and improves the stability of the training process.

D. Centralized Training and Decentralized Implementation

The implementation of the proposed approach can be divided into two main steps: centralized training for the formulation of coordinated strategies and decentralized execution for voltage regulation with only local information. They are explained below.

1) Centralized Training

In the MGs with *M* agents, the parameter set to be optimized is $\mathrm{Q} = \{\mathrm{Q}_1, ..., \mathrm{Q}_M\}$. For agent *j*, the parameter set is denoted as $\mathrm{Q}_j = \{\mathrm{Q}^{\mu_j}, \mathrm{Q}^{\mu'_j}, \mathrm{Q}^{Q_j}, \mathrm{Q}^{Q'_j}\}$, where Q^{μ_j} and $\mathrm{Q}^{\mu'_j}$ are parameters of actor and target actor network of agent *j*; Q^{Q_j} and $\mathrm{Q}^{Q'_j}$ are parameters of attention-critic and target attention-critic network. The training process of the proposed approach is shown in Algorithm I.

The parameters of NN start to update when the replay buffer is full. At each time-step, each agent samples a mini-batch of experiences $(\mathbf{s}_j, \mathbf{a}_j, r, \mathbf{s}'_j)_k$, $k = 1,2,...,B$ from its memory. Each actor NN takes the local state $\mathbf{s}_j$ as input, and adjust its parameters to output an action that maximizes the action value. The gradient is calculated according to

$$\nabla_{\mathrm{Q}^{\mu_j}} J(\mathrm{Q}^{\mu_j}) = \frac{1}{B}\sum_{k=1}^{B} \nabla_{\mathrm{Q}^{\mu_j}} p_j(\mathbf{a}_j | \mathbf{s}_j) \nabla_{\mathbf{a}_j} Q_j(f_j(\mathbf{s}_j, \mathbf{a}_j), e_j)|_{\mathbf{a}_j = p_j(\mathbf{s}_j)} \quad (21)$$

Then, the parameters of actor network are updated through

$$\mathrm{Q}^{\mu_j} \leftarrow \mathrm{Q}^{\mu_j} + \eta_\mu \nabla_{\mathrm{Q}^{\mu_j}} J(\mathrm{Q}^{\mu_j}) \quad (22)$$

where η_μ is the learning rate for actor function. Each critic NN takes the global state as input, which includes the states and actions of all agents, and predicts an action value to minimize the following loss:

$$L(\mathrm{Q}^{Q_j}) = \frac{1}{B}\sum_{k=1}^{B} (Q_{j,n}(f_j(\mathbf{s}_j, \mathbf{a}_j), e_j) - y)^2, \ n = 1,2 \quad (23)$$

Algorithm I Training Process of the Proposed Algorithm

Algorithm Training the proposed algorithm

Input: the node active and reactive power in the DN, the PV output, and the reward.

Output: DNN's parameters Q

Randomly initialize parameters of critic networks Q^{Q_j} and actor network Q^{μ_j} for each agent *j*

Initialize target networks $\mathrm{Q}^{Q'_j} \leftarrow \mathrm{Q}^{Q_j}, \mathrm{Q}^{\mu'_j} \leftarrow \mathrm{Q}^{\mu_j}$ for each agent *j*

for episode =1, 2, …, *H*

  receive initial observation $\mathbf{s}_j$ for each agent *j*

  for *t*=1, 2, …*T*

    choose action according to $\mathbf{a}_j = p_j(\mathbf{s}_j)$ for each agent *j*

    execute actions $\mathbf{a} = (\mathbf{a}_1, ..., \mathbf{a}_M)$ and obtain reward *r* and new observation $\mathbf{s}'_j$ for each agent *j*

    store transition $(\mathbf{s}_j, \mathbf{a}_j, r, \mathbf{s}'_j)$ in the replay buffer

    for agent *j* = 1,…, *M*

      sample a random mini-batch *B* of transitions from replay buffer

      calculate target *y* according to (17)

      update critic networks according to (23) and (24)

      if *t* mod *d*

        update actor network according to (21) and (22)

        update target networks according to (25)

      end if

    end for

  end for

end for

The parameters are updated according to

$$Q^{Q_j} \leftarrow Q^{Q_j} + \eta_Q \nabla_{Q^{Q_j}} L(Q^{Q_j}) \tag{24}$$

where η_Q is the learning rate for the critic function. Then, the parameters of target networks are optimized by slowly tracking the online ones:

$$Q^{Q_j'} \leftarrow \tau Q^{Q_j} + (1-\tau) Q^{Q_j'}, Q^{\mu_j'} \leftarrow \tau Q^{\mu_j} + (1-\tau) Q^{\mu_j'} \tag{25}$$

where τ <<1 is the soft update coefficient.

During the training process, the critic of each agent requires information from other agents. This can be achieved since the training of DRL is implemented offline. The centralized critics that are augmented with information about other agents' policies during training help identify coordinated strategies. The explicitly modeling of other agents' decision-making process allows each agent to provide decisions with better robustness to system dynamics based on local information only. This differentiates the existing works and allows us to deal with scalability issues in the presence of large-scale systems.

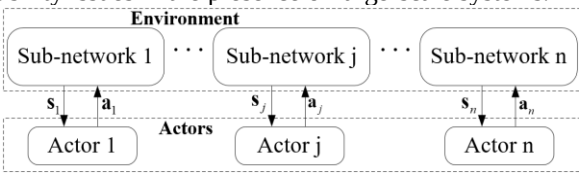


Fig. 2. The workflow of the decentralized execution.

2) Decentralized Execution

When the training process is completed, the parameters of DNN are fixed and only the actor network of each agent is kept for real-time voltage regulation. The workflow of decentralized execution is shown in Fig. 2. Each agent is in charge of a sub-network. Since the actors only need local information, the proposed approach can be executed in a decentralized manner. The real-time reactive power control scheme of the proposed approach is shown in Algorithm II.

Algorithm II Real-time Reactive Power Control of the Proposed Approach

Algorithm Real-time reactive power control

Input: the node active and reactive power in the DN, the PV output.

Output: reactive power schedules $\mathbf{a}$.

1: Read the parameters of actor network of each agent Q^{μ_j}

2: For time step t=1, 2, …T

3: for agent j = 1, …, M

4: obtain the local observation $\mathbf{s}_j$

5: calculate action $\mathbf{a}_j$ according to $\mathbf{a}_j = p_j(\mathbf{s}_j)$

6: end for

7: concatenate actions of all agents $\mathbf{a} = (\mathbf{a}_1, ..., \mathbf{a}_M)$

8: end for

IV. Numerical Results

In this section, simulation results are provided to evaluate the performance of the proposed approach on IEEE 33-bus and 123-bus systems, whose parameters can be found in [38]. The network partition results are first illustrated followed by the control performance comparison results with other methods.

A. Simulation Setup

To simulate more realistic scenarios, real-world PV data are used, i.e., one-year PV generation data of Xiaojin, a county in the Sichuan province of China. These data are divided into a training set and a test set, which contain 300- and 10-days' data, respectively. The sampling frequency of the data is one hour. Note that the strategy learned by the proposed approach can be easily extended to scenarios with different sampling times in practice. The parameters of the control devices are shown in Table I. The maximum voltage deviation is set to $\pm 5\%$. For the proposed method, each sub-region is modeled as an agent, which is composed of actor and critic networks. All the networks have two shallow layers, the number of neurons of

Table I Parameters of Control Devices for 33-bus System

Type	Capacity	Location
SVC	0.3MVar	5, 10, 30
PV	0.8MW/0.8MVA	15, 18, 22, 24, 27, 33

Table II Parameter Settings of the Proposed Method

Parameters	Values
Batch size for updating NN	32
Replay buffer size	48000
Discount factor	0
Soft update coefficient	0.001
Policy update frequency	2
Target policy smoothing coefficient	0.2
Learning rate for actor network	0.001
Learning rate for critic network	0.002

which are 100 and 100, respectively. The hyper-parameters are shown in Table II. The power flow is carried out using Matlab and the training of the proposed method is implemented in Python with TensorFlow. A workstation with an NVIDIA GeForce 1080Ti GPU and an Intel Xeon E5-2630 CPU is used for the simulation.

TABLE III
Voltage Deviation of Various Methods on 33-Bus System

Cluster number	3	4	5	6
Ave. vol. dev. of MADDPG [30]	0.13%	0.13%	0.14%	0.17%
Ave. vol. dev. of proposed method	0.13%	0.13%	0.13%	0.13%

B. Sensitivity of Volt-VAR Control with Network Partition

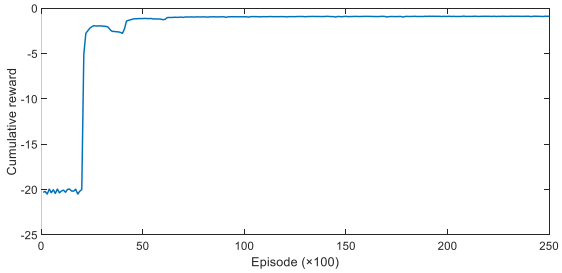


Fig. 3. The evolution of the reward during the training procedure.

The spectral clustering algorithm is applied to partition the ADN into several sub-regions. After that, the proposed approach is trained for 25000 epochs on the training data to learn the coordinated control strategy for voltage regulation. Each epoch corresponds to a day, which is randomly selected from the training set at the beginning of each epoch. The

convergence curve of the cumulative reward is plotted in Fig. 3. It can be observed that the proposed approach could not make balanced decisions at the beginning of the training procedure and therefore achieves low reward. With the training process going on, the reward increases significantly and finally converges around -0.96 with small fluctuations. This illustrates that the proposed method can learn the coordinated control strategy from training data.

After training, the average voltage deviations on test data for MADDPG [30] and the proposed approach under different partition results are obtained and shown in Table III. The MADDPG method employs the centralized training and decentralized execution framework and models each sub-network as a DDPG agent. The network structure and hyper-parameter settings are the same as our proposed method. *It is worth noting that the MADDPG method was applied to the transmission network in [30] and we customize it for the voltage regulation of the distribution system for a fair comparison.* It can be observed from the table that the control performance of the MADDPG method decreases when the number of sub-regions is gradually increased. This is because the complexity of the environment increases with the growing number of agents. This makes it difficult for the MADDPG method to search for a good policy. However, the proposed method can maintain control performance with an increased number of agents thanks to the developed attention model. The latter helps each agent attend to specific information that is mostly related to the reward. The partition results may be different according to the actual requirement of operators and our proposed method can easily adapt to them.

In the subsequent tests, we select the number of sub-networks as 6 in the 33-bus system for illustrations. The clustering result is shown in Fig. 4.

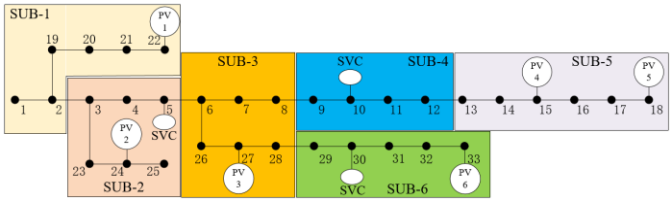


Fig. 4. The partition results of the IEEE 33-bus system.

C. Comparison Results with Other Alternatives

To demonstrate the benefits of the proposed method, comparative tests are carried out against various existing methods. They include 1) **the no control method**; 2) **the TD3-D method**, where each sub-network is controlled by a TD3 agent based on the local observation. The TD3 agents are trained separately and sequentially to minimize the voltage deviation of their sub-network. Note that there is no information exchange between TD3 agents during the training process; 3) **the stochastic programming-based (SP) approach**, where all the sub-networks are optimized separately, and the objective of each sub-network is to minimize the voltage deviation based on local information. 300 scenarios are randomly generated to represent the uncertainty and scenarios reduction is used to obtain 20 representative scenarios [7]; 4) **the model-based centralized control method**, where load demand and PV generation are assumed to be known beforehand and the commercial SOCP solver MOSEK is applied to solve deterministic cases based on the global information. Its results with the *perfect model* are considered as benchmarks.

The average, maximum rise, and maximum drop of voltage deviations, as well as the computing times for all methods, are shown in Table IV. It can be observed that when reactive power compensation is not applied, the voltage will exceed the upper and lower limits. The TD3-D, MADDPG, and the proposed approach can ensure the voltage to be within the limited ranges. Note that these three methods all make decisions based on local information. However, the MADDPG method achieves better performance due to the coordinated control strategies learned during the training process. The proposed approach further enhances the control performance thanks to the use of the attention mechanism. The proposed method achieves similar performance with that obtained by the SP method, which adopts the centralized control framework and informs decisions based on global information. To evaluate the control accuracy of the proposed method, the average optimization error is defined:

TABLE IV Voltage Deviation of Various Methods

Method	Average	Max rise	Max drop	Com. type
No control	1.46%	5.22%	7.11%	-
TD3-D	0.77%	3.88%	2.03%	D
MADDPG [30]	0.17%	1.32%	1.25%	D
SP [7]	0.13%	1.69%	1.24%	C
Proposed	0.13%	0.73%	1.25%	D
Centralized	0.08%	0.55%	1.25%	C

$$ERR=\left|\frac{\Delta V_{pro}-\Delta V_{cen}}{\Delta V_{cen}-\Delta V_{ori}}\right|\times 100\% \tag{26}$$

where *ERR* represents the average deviation of the proposed approach to the global optimal solution; ΔV_{cen}, ΔV_{pro} and ΔV_{ori} represent the average voltage deviations of the centralized approach, the proposed approach, and the original value on test data, respectively. The *ERR* is 3.6% for the proposed approach and this means that it can reach 96.4% optimality based on local information. However, it assumes that the uncertain variables to be known beforehand, which is impossible to obtain in practice. Also, this method depends on complete two-way communication links. The infrastructure for this type of control framework is rarely available in ADNs. Note that the proposed approach is based on the decentralized control framework, where each sub-region is controlled using only local information without the communications between agents. Thus the communication requirements are reduced and the privacy concerns are mitigated.

To further demonstrate the capability of the proposed method in dealing with fluctuations of PV and load outputs, a sunny day is selected as a case study. The PV output and load demand for a sunny day are shown in Fig. 5 and Fig. 6, respectively. The voltage profile of each node obtained by various control strategies is shown in Fig. 7 when *t=1:00 PM*. It can be observed that the voltages at nodes 17-18 go beyond the upper voltage limit when there is no reactive compensation. With the TD3-D method, the voltages can be adjusted to a limited range. However, due to the lack of coordination of multi-devices, the capability of the voltage regulation devices is not fully utilized. Since the MADDPG method and the proposed approach explicitly model the decision-making process of other agents via centralized training, the agents can exhibit cooperative

behaviors with only local information. The proposed approach further enhances the control performance of MADDPG via the attention model. The voltage profile of the proposed approach is very close to that of the SP and centralized method, demonstrating the effectiveness of the proposed method.

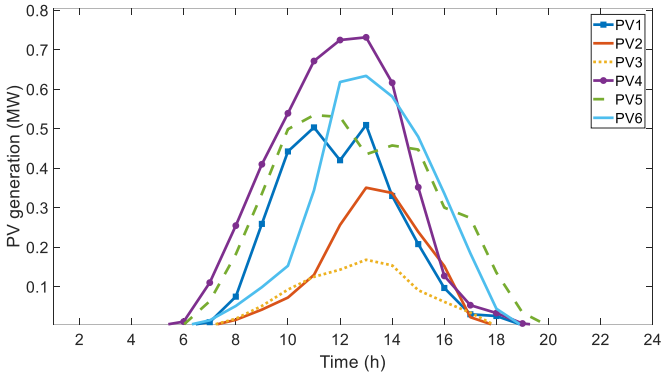


Fig. 5. PV outputs for the selected sunny day.

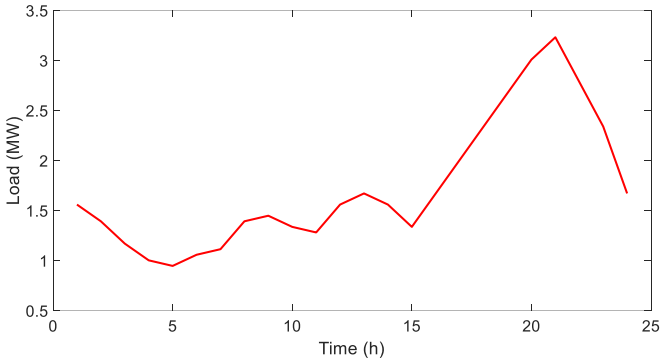


Fig. 6. Load demand for the selected sunny day.

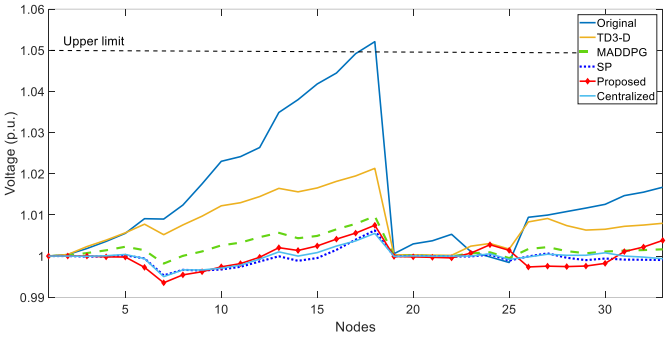


Fig. 7. Voltages of each node before and after control when *t=1:00* PM.

D. Robustness to Large Stochasticity

More simulations are carried out to demonstrate the advantage of the proposed approach for real-time controls. The PV output profile is shown in Fig. 8. Due to the cloud dynamics, the PV output changes fast within 1 minute, i.e., its output rises from 0.33 MW at *t=1s*, reaches 0.65 MW at *t=30s*, and then returns to 0.33 MW in 30 seconds. The voltage profiles of node 18 under various control strategies are shown in Fig. 9. Since the SP method suffers from a heavy computational burden, it provides a predetermined control solution to deal with the rapid PV output variation during the short period. The TD3-D, MADDPG, and the proposed methods are DRL based approaches, the strategy learned by which can generalize to new situations and inform decisions in milliseconds according to the latest observation. In this test, they provide decisions at each second. The centralized method ignores the communication delay and provides a theoretical limit for the problem (an ideal condition that could not be achieved in practice). It can be observed from Fig. 8 that node 18 suffers from an overvoltage problem if no controls of SVC and PV inverter are applied, namely the original method. With the SP method, the problem can be suppressed. However, since the control decisions provided by this approach are predetermined and cannot react dynamically to the fast-changing PV outputs, it has much larger voltage fluctuations than those of the TD3-D, MADDPG, and the proposed method. By contrast, the TD3-D, MADDPG, and the proposed approach can make decisions based on the latest states of the ADN, and thus can achieve a better response to the dynamic changes of the PV outputs. The proposed approach outperforms TD3-D and MADDPG due to the coordinated control strategy learned during the training process and the utilization of the attention model.

TABLE V Control Devices in 123-bus System

Type	Capacity	Location
SVC	0.3MVar	9, 35, 54,62,68,105
PV	0.8MW/0.84MVA	5,12,27,50,65,76,81,83,100, 114,118

TABLE VI Control Performance on IEEE 123-bus System

Cluster number	6	7	8
Ave. vol. dev. of MADDPG [30]	0.60%	0.63%	0.80%
Ave. vol. dev. of proposed method	0.45%	0.45%	0.46%

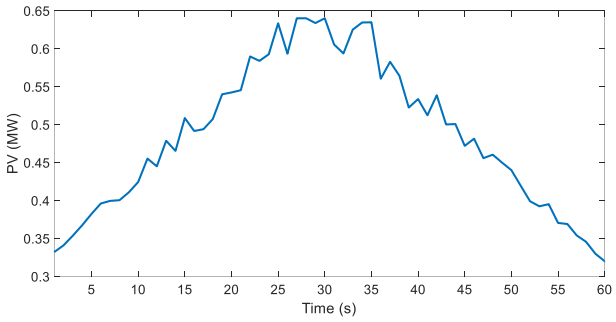


Fig. 8. PV outputs with large variations.

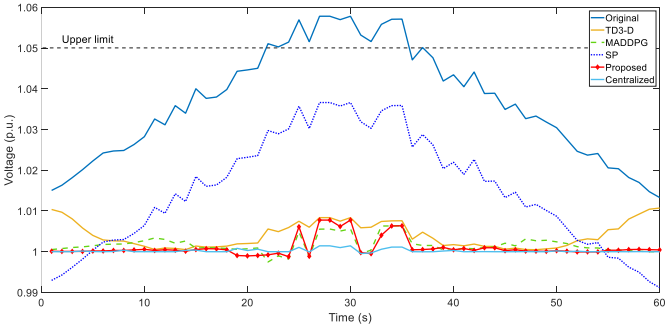


Fig. 9. Voltage change of node 18 with different control strategies when the PV outputs have large fluctuations.

E. Scalability to IEEE 123-bus System

To assess the scalability of the proposed method to a larger-scale system, tests are carried out on the IEEE 123-bus system [38]. The parameter settings of the control devices are shown in Table V, including the capacities and locations of SVCs and PVs. The average voltage deviations on test data achieved by

the MADDPG and the proposed approach under different network partition results are listed in Table VI. The parameter settings of the proposed control method are the same as those for the IEEE 33-bus system. The MADDPG method suffers from performance degradation when the number of sub-regions increases. This is not the case for the proposed method. The conclusion is consistent with that in the 33-bus system. The number of sub-regions is set as 8 in this paper. The partition result is shown in Fig. 10.

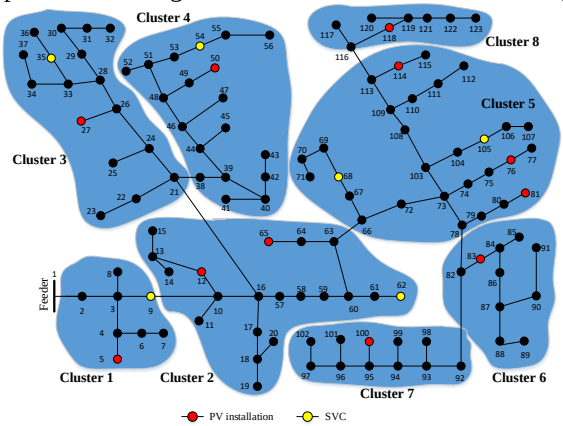


Fig. 10. The partition results of the IEEE 123-bus system.

TABLE VII Voltage Deviations for IEEE 123-bus System

Method	Average	Max rise	Max drop	Com. type
No control	1.74%	6.21%	4.33%	-
TD3-D	2.94%	1.91%	8.65%	C
MADDPG [30]	0.80%	2.40%	3.01%	D
SP [7]	0.42%	1.40%	3.31%	C
Proposed	0.46%	2.08%	2.79%	D
Centralized	0.32%	1.11%	2.90%	C

The voltage deviations obtained from different approaches are shown in Table VII. It can be found that if there is no control, the maximum voltage rise will be 6.21%. It is very interesting to find that the TD3-D method has serious voltage control issues for the 123-bus system and fails to find a good voltage regulation strategy based on local information. With the increased size of the system and the problem complexity, the negative impacts of not coordinating with different sub-networks have been shown here. Both the MADDPG method and the proposed approach are implemented with centralized training and distributed execution. With the MADDPG method, voltages can be adjusted to a limited range. However, as compared to the proposed approach, the MADDPG method has a larger voltage fluctuation due to the immense growth of the complexity of the environment by the increased number of agents. The proposed approach can achieve better results with the attention mechanism.

V. Conclusions

This paper proposes a distributed coordination control for distribution system Volt-VAR control considering PV inverters and SVCs. The spectral clustering algorithm allows us to partition the large distribution system into several sub-networks from the voltage control perspective. Then, the control of each sub-network is formulated as the MGs and solved by the attention-based MATD3 algorithm. The proposed method is centralized training distributed implementation and can be easily used for real-time voltage regulation. Compared with centralized control, the proposed approach mitigates the issues, such as the communication bottleneck and privacy concerns. Compared with other distributed control methods, only local information is needed without communications between agents. The proposed method can adapt to the flexible network partition requirements of the operator than the typical MADRL algorithm. Comparative results with several other existing model-based and data-driven methods demonstrate that the proposed method can achieve 96.4% optimality based on local information while considering the uncertainty. However, the model-based could not achieve satisfactory outcomes in the presence of rapid variations of PV outputs. The future works include the development of a new control method that can coordinate the smart inverters and utility-owned equipment, which is a two-timescale control problem. We will also propose a meta-learning based MADRL algorithm to deal with topology changes in the distribution networks.

References

[1] R. Detchon, R. Van Leeuwen, “Policy: bring sustainable energy to the developing world.” *Nature*, vol. 508, no. 7496, pp. 309-311, 2014.

[2] V. Quezada, J. Abbad, T. Roman, “Assessment of energy distribution losses for increasing penetration of distributed generation.” *IEEE Trans. Power Syst.*, vol. 21, no. 2, pp. 533-540, 2006.

[3] Y. J. Kim, J. L. Kirtley, and L. K. Norford, “Reactive power ancillary service of synchronous DGs in coordination with voltage control devices,” *IEEE Trans. Smart Grid*, vol. 8, no. 2, pp. 515–527, Mar. 2017.

[4] B. Stott, J. Jardim, and O. Alsac, “DC power flow revisited,” *IEEE Trans. Power Syst.*, vol. 24, no. 3, pp. 1290–1300, 2009.

[5] W. Min, S. Liu, “A trust region interior point algorithm for optimal power flow problems,” *Int. J. Electr. Power Energy Syst.*, vol. 2, pp. 293–300, 2005.

[6] D. K. Molzahn, I. A. Hiskens, "Sparsity-exploiting moment-based relaxations of the optimal power flow problem," *IEEE Trans. Power Syst.*, vol. 30, no. 6, pp. 3168-3180, 2015..

[7] Y. Xu, Z. Y. Dong, R. Zhang, *et al.*, "Multi-timescale coordinated voltage/var control of high renewable-penetrated distribution systems," *IEEE Trans. Power Syst.*, vol. 32, no. 6, pp. 4398-4408, Nov. 2017.

[8] T. Ding, C. Li, Y. Yang, *et al.*, "A two-stage robust optimization for centralized-optimal dispatch of photovoltaic inverters in active distribution networks," *IEEE Trans. Sustain. Energy*, vol. 8, no. 2, pp. 744-754, Apr. 2017.

[9] T. Soares, R. J. Bessa, P. Pinson and H. Morais, "Active distribution grid management based on robust AC optimal power flow," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6229-6241, Nov. 2018.

[10] P. Li, C. Zhang, Z. Wu, *et al.*, "Distributed adaptive robust voltage/var control with network partition in active distribution networks." *IEEE Trans. Smart Grid*, 2019.

[11] Y. Guo, Q. Wu, H. Gao, *et al.* "MPC-based coordinated voltage regulation for distribution networks with distributed generation and energy storage system." *IEEE Trans. Sustain. Energy*, vol. 10, no. 4, pp. 1731-1739, Oct. 2019.

[12] Z. Wang, J. Wang, B. Chen, M. Begovic, and Y. He, “MPC-based voltage/var optimization for distribution circuits with distributed generators and exponential load models,” *IEEE Trans. Smart Grid*, vol. 5, no. 5, pp. 2412-2420, Sep. 2014.

[13] X. Liu, A. Aichhorn, L. Liu, and H. Li, “Coordinated control of distributed energy storage system with tap changer transformers for voltage rise mitigation under high photovoltaic penetration,” *IEEE Trans. Smart Grid*, vol. 3, no. 2, pp. 897–906, Jun. 2012.

[14] G. Cavraro and R. Carli, “Local and distributed voltage control algorithms in distribution network,” *IEEE Trans. Power Syst.*, vol. 33, no. 2, pp. 1420–1430, Mar. 2018.

[15] H. Xin, Y. Liu, Z. Qu, and D. Gan, “Distributed control and generation estimation method for integrating high-density photovoltaic systems,” *IEEE Trans. Energy Conversion*, vol. 29, no. 4, pp. 988–996, Dec. 2014.

[16] M. Zeraati, M. Golshan, J. Guerrero, "Distributed control of battery energy storage systems for voltage regulation in distribution networks with high PV penetration." *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 3582-3593, Jul. 2018.

[17] Y. Chai, L. Guo, C. Wang, *et al*., "Network partition and voltage coordination control for distribution networks with high penetration of distributed PV units," *IEEE Trans. Power Syst.*, vol. 33, no. 3, pp. 3396-3407, May 2018.

[18] B. Zhao, Z. Xu, C. Xu, *et al*., "Network partition-based zonal voltage control for distribution networks with distributed PV systems." *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 4087-4098, Sep. 2018.

[19] W. Zheng, W. Wu, B. Zhang, H. Sun, Y. Liu, "A fully distributed reactive power optimization and control method for active distribution networks", *IEEE Trans. Smart Grid*, vol. 7, no. 2, pp. 1021-1033, 2016.

[20] G. Wang, V. Kekatos, A. J. Conejo, and G. B. Giannakis, "Ergodic energy management leveraging resource variability in distribution grids," *IEEE Trans. Power Syst.*, vol. 31, no. 6, pp. 4765–4775, Nov. 2016.

[21] Wang, Wei, Nanpeng Yu, Yuanqi Gao, and Jie Shi, "Safe off-policy deep reinforcement learning algorithm for volt-var control in power distribution systems." *IEEE Trans. Smart Grid*, 2019.

[22] D. Cao, *et al*., "Reinforcement learning and its applications in modern power and energy systems: a review." *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 6, pp. 1029-1042, 2020.

[23] Mocanu, E., Mocanu, D. C., Nguyen, P. H., Liotta, A., Webber, M. E., Gibescu, M., & Slootweg, J. G., "On-line building energy optimization using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3698 - 3708, Jul. 2019.

[24] V. Bui, A. Hussain and H. Kim, "Double deep Q-learning-based distributed operation of battery energy storage system considering uncertainties," *IEEE Trans. Smart Grid*, vol. 11, no. 1, pp. 457-469, 2019.

[25] Q. Yang, G. Wang, A. Sadeghi, G. B. Giannakis and J. Sun, "Two-timescale voltage control in distribution grids using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2313-2323, May 2020.

[26] G. Zhang, *et al*., "Deep reinforcement learning-based approach for proportional resonance power system stabilizer to prevent ultra-low-frequency oscillations," *IEEE Trans. Smart Grid*, vol. 11, no. 6, pp. 5260-5272, Nov. 2020.

[27] Y. Xu, W. Zhang, W. Liu and F. Ferrese, "Multiagent-based reinforcement learning for optimal reactive power dispatch," *IEEE Trans. Systems, Man, and Cybernetics, Part C* (Applications and Reviews), vol. 42, no. 6, pp. 1742-1751, Nov. 2012.

[28] F. Zhu, Z. Yang, F. Lin and Y. Xin, "Decentralized cooperative control of multiple energy storage systems in urban railway based on multiagent deep reinforcement learning," *IEEE Trans. Power Elec*., vol. 35, no. 9, pp. 9368-9379, Sept. 2020.

[29] Y. Zhang, *et al*., "Deep reinforcement learning based volt-VAR optimization in smart distribution systems," *IEEE Trans. Smart Grid*, vol. 12, no. 1, pp. 361-371, Jan. 2021.

[30] S. Wang, *et al*., "A data-driven multi-agent autonomous voltage control framework using deep reinforcement learning," *IEEE Trans. Power Syst*., vol. 35, no. 6, pp. 4644-4654, Nov. 2020.

[31] H. Zhu and H. J. Liu, "Fast local voltage control under limited reactive power: optimality and stability analysis," *IEEE Trans. Power Syst*., vol. 31, no. 5, pp. 3794-3803, Sept. 2016.

[32] S. Kaisler, F. Armour, and J. A. Espinosa, "Introduction to big data: Challenges, opportunities, and realities minitrack," in *Proc. 47th Hawaii Int. Conf. Syst. Sci. (HICSS)*, Waikoloa, HI, USA, 2014, p. 728.

[33] S. M. Amin, "Smart grid security, privacy, and resilient architectures: opportunities and challenges," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, San Diego, CA, USA, 2012, pp. 1–2.

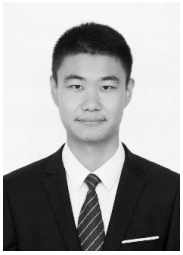
[34] A.Y. Ng, M.I. Jordan, and Y Weiss. "On spectral clustering: analysis and an algorithm." Adv Neural Inf Proc Syst, MIT Press, 2001.

[35] I. Shariq, F. Sha, "Actor-attention-critic for multi-agent reinforcement learning," International Conference on Machine Learning, Long Beach, CA, USA, June, 2019.

[36] R. Lowe, Y. Wu, A. Tamar, *et al.* "Multi-agent actor-critic for mixed cooperative-competitive environments." *Advances in Neural Information Processing Systems*, pp. 6379-6390, 2017.

[37] S. Fujimoto, H. Van Hoof, D. Meger. "Addressing function approximation error in actor-critic methods." *arXiv preprint arXiv*:1802.09477, 2018.

[38] IEEE PES, Distribution Test Feeders, Sep. 2010. [Online]. Available: https://site.ieee.org/pes-testfeeders/resources/.



Di Cao is currently working toward the Ph.D. degree in control science and engineering at the University of Electronic Science and Technology of China. His research interest includes optimization of distribution network and applications of machine learning in power systems..

Junbo Zhao (SM'19) received the Ph.D. degree in Electrical Engineering from the Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, USA in 2018. He was a Research Assistant Professor at Virginia Tech from May 2018 to August 2019. He did the summer internship at Pacific Northwest National Laboratory from May to August 2017. He is currently an Assistant Professor with Mississippi State University, Starkville, MS, USA. His research interests include cyber-physical power system modeling estimation security dynamics and stability, uncertainty quantification, robust statistical signal processing, and machine learning for smart grids. He serves as the editor of IEEE Transactions on Power Systems, IEEE Transactions on Smart Grid and IEEE Power and Engineering Letters, the Associate Editor of International Journal of Electrical Power Energy Systems, and the subject editor of IET Generation, Transmission & Distribution.

Weihao Hu (S'06–M'13–SM'15) received the B.Eng. and M.Sc. degrees from Xi'an Jiaotong University, Xi'an, China, in 2004 and 2007, respectively, both in electrical engineering, and Ph. D. degree from Aalborg University, Denmark, in 2012.

He is currently a Full Professor and the Director of Institute of Smart Power and Energy Systems (ISPES) at the University of Electronics Science and Technology of China (UESTC). He was an Associate Professor at the Department of Energy Technology, Aalborg University, Denmark and the Vice Program Leader of Wind Power System Research Program at the same department. His research interests include artificial intelligence in modern power systems and renewable power generation. He has led/participated in more than 15 national and international research projects and he has more than 170 publications in his technical field.

He is an Associate Editor for IET Renewable Power Generation, a Guest Editor-in-Chief for Journal of Modern Power Systems and Clean Energy Special Issue on Applications of Artificial Intelligence in Modern Power Systems, a Guest Editor-in-Chief for Transactions of China Electrical Technology Special Issue on Planning and operation of multiple renewable energy complementary power generation systems, and a Guest Editor for the IEEE TRANSACTIONS ON POWER SYSTEM Special Section on Enabling very high penetration renewable energy integration into future power systems. He was serving as the Technical Program Chair (TPC) for IEEE Innovative Smart Grid Technologies (ISGT) Asia 2019 and is serving as the Conference Chair for the Asia Energy and Electrical Engineering Symposium (AEEES 2020). He is currently serving as Chair for IEEE Chengdu Section PELS Chapter. He is a Fellow of the Institution of Engineering and Technology, London, U.K. and an IEEE Senior Member.

Fei Ding (S'12–M'14–SM'18) received her Ph.D. from Case Western Reserve University, and she joined the National Renewable Energy Laboratory as a research engineer in 2015. Her research focuses on distribution system automation and optimization, distribution system modeling and simulation, renewable energy grid integration, advanced distribution management system, smart grid resilience.

Qi Huang (S'99, M'03, SM'09) was born in Guizhou province in the People's Republic of China. He received his BS degree in Electrical Engineering from Fuzhou University in 1996, MS degree from Tsinghua University in 1999, and Ph.D. degree from Arizona State University in 2003. He is currently a professor at UESTC, the Executive Dean of School of Energy Science and Engineering, UESTC, and the director of Sichuan State Provincial Lab of Power System Wide-area Measurement and Control. He is a member of IEEE since 1999. His current research and academic interests include power system instrumentation, power system monitoring and control, and power system high performance computing.

Zhe Chen (M'95-SM'98-F'19) received the B.Eng. and M.Sc. degrees from Northeast China Institute of Electric Power Engineering, Jilin, China, and the Ph.D. degree from University of Durham, Durham, U.K.

He is a Full Professor with the Department of Energy Technology, Aalborg University, Denmark. He is the Leader of Wind Power System Research Program in the Department of Energy Technology, Aalborg University and the Danish Principle Investigator for Wind Energy of Sino-Danish Centre for Education and Research. His research areas include power systems, power electronics and electric machines; and his main current research interests are wind energy and modern power systems. He has led many research projects and has more than 400 publications in his technical field.